

The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation

Barbara Plank

Center for Information and Language Processing (CIS), MaiNLP lab, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
b.plank@lmu.de

Abstract

Human variation in labeling is often considered noise. Annotation projects for machine learning (ML) aim at minimizing human label variation, with the assumption to maximize data quality and in turn optimize and maximize machine learning metrics. However, this conventional practice assumes that there exists a *ground truth*, and neglects that there exists genuine human variation in labeling due to disagreement, subjectivity in annotation or multiple plausible answers. In this position paper, we argue that this big open problem of *human label variation* persists and critically needs more attention to move our field forward. This is because human label variation impacts all stages of the ML pipeline: *data, modeling and evaluation*. However, few works consider all of these dimensions jointly; and existing research is fragmented. We reconcile different previously proposed notions of human label variation, provide a repository of publicly-available datasets with un-aggregated labels, depict approaches proposed so far, identify gaps and suggest ways forward. As datasets are becoming increasingly available, we hope that this synthesized view on the “problem” will lead to an open discussion on possible strategies to devise fundamentally new directions.

1 Introduction

In Natural Language Processing (NLP) much progress today is driven by fine-tuning large pre-trained language models using an annotated dataset, assumed to be representative for a target language task of interest (Schlangen, 2021). This is analogously so in Machine Learning (ML) and Computer Vision (CV), where the target tasks differ, yet the conceptual pipeline remains the same: data, modeling, evaluation. Despite the *importance of annotated data*—as it fuels all steps in this pipeline—a crucial assumption of today’s learning systems is to rely on a single gold label per instance. The gold

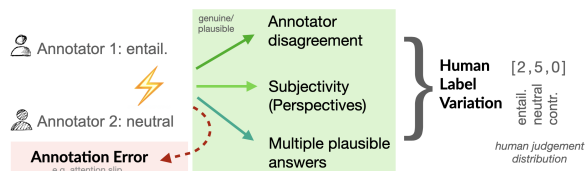


Figure 1: We propose the term *human label variation* to capture the fact that inherent disagreement in annotation can be due to genuine disagreement, subjectivity or simply because two (or more) views are plausible.

label is obtained by aggregation (e.g. majority vote) of labels crucially provided by *humans*.

The assumption of a *ground truth* (and taking the majority vote or the ‘mode’ of the human judgement distribution) makes sense when humans involved in labeling highly agree on the answer to the questions, such as “Does this image contain a bird?”, “Is ‘learn’ a verb?”, “What is the capital of Italy?”. However, this assumption often does not make sense—especially when language is involved. For example, on questions determining a word sense, questions such as “Is this comment toxic?” or questions involving understanding indirect answers to polar questions like “Q: Hey. Everything ok?” “A: I’m just mad at my agent” (see more examples in Figure 2). While some disagreement is due to human labeling errors (cf. Figure 1 arrow to the left and § 3), an increasing body of work has shown that irreconcilable variation between annotations is plausible and abundant (Plank et al., 2014b; Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019; Uma et al., 2021b) (illustrated in Figure 1). The observed variation can indeed be disagreement due to difficult cases, subjectivity or cases where multiple answers are plausible (cf. § 2). We argue that human label variation (HLV) provides rich information that should not be discarded. Critically, to rely on a ground truth means we tacitly agree to continue: i) to create datasets that encode a single

ground truth, ii) to develop models that are optimized towards a single preferred output, and iii) to evaluate models against a single ground truth. By continuing to do so, we might ask ourselves if we are climbing the right hill—or whether continuing to model a single ground truth hampers progress.

In this position paper, we argue that neglecting variation in labeling is problematic, as it impacts *all steps of the pipeline*. Traditionally, this variation has been considered a problem. We underline emerging works that instead believe this issue to be an opportunity. In fact, we believe it is essential to take human label variation into account for progress. Human labels are bound to be scarce yet at the same time critical as they provide human interpretations *and values*. Therefore, embracing it is necessary for human-facing NLP, i.e., technology which is by and for humans; inclusive and reliable. However, the research landscape is fragmented, and approaches often focus on either steps of the pipeline. Therefore, in this paper we focus on the three core aspects of the pipeline: data, modeling and evaluation. In particular, i) we distill some of the on-going discussions in disparate (sub-)fields and propose a unified term; ii) we present and work out suggestions for each for the future; and iii) we provide a comprehensive repository of publicly-available data sets that allow studying human label variation, and invite the community to contribute.

2 Data and Human Label Variation

High-quality data is essential for any empirical scientific inquiry and has to satisfy the requirements of validity and reliability (Krippendorff, 2018; Pustejovsky and Stubbs, 2012; Schlangen, 2021). However, for almost all tasks in NLP and CV irreconcilable disagreement between annotators has been observed (Uma et al., 2021b). In light of this, the original definition of data reliability is questionable—it assumes labels follow a given standard. We might ask which standard?

Human annotations are needed to ground and make sense of language, images, speech etc. However, labelling data is difficult, particularly when dealing with an object of study as complex as language. Take the illustration in Figure 2 as example. While categories exist, their boundaries are fluid, or simply multiple options are plausible.

Disagreement or variation? We define *human label variation* (HLV) as plausible variation in annotation, see Figure 1, to reconcile different no-


<i>Positive, Negative or Neutral?</i> This was an excellent offer and i didn't not enjoy it.	<i>Entailment: does A entail B?</i> A : Paola drinks oat milk. B : Paola drinks milk.	<i>Image classification: dog, cat,...?</i> 
-------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------

Figure 2: Hard cases. Image from (Uma et al., 2021b).

tions found in the literature (discussed next). We prefer ‘variation’, because ‘*disagreement*’ implies that two (or more) views involved cannot all hold. In contrast, errors are annotation differences, due to amongst others attention slips. Crucially, HLV assumes humans usually provide their best judgements, and variation emerges due to, e.g., ambiguity of the instance, uncertainty of the annotator, genuine disagreement, or simply the fact that multiple options are correct. Aggregation obfuscates this real-world complexity.

HLV has been studied in CV, where it is dubbed *human uncertainty* (Peterson et al., 2019), as well as in human-computer interaction (HCI) as *disagreement* or *contested labels* (Gordon et al., 2021). In NLP, variation has been acknowledged as *annotator disagreement* already in early works on resolving disagreement (Poesio and Artstein, 2005), particularly in pragmatics and discourse (de Marneffe et al., 2012; Webber and Joshi, 2012; Das et al., 2017). HLV in NLP is discussed both from the **linguistic side** as *hard cases* (Zeman, 2010; Plank et al., 2014b), *difficult linguistic cases* (Manning, 2011), as judgements which are *not always categorical* (de Marneffe et al., 2012), *inherent disagreement* (Pavlick and Kwiatkowski, 2019; Davani et al., 2022) and *justified and informative disagreement* (Sommerauer et al., 2020). Variation in NLP is also discussed in connection to **subjectivity**, e.g., as *a range of reasonable interpretations* (CrowdTruth) (Aroyo and Welty, 2015), as *one or many beliefs* (Rottger et al., 2022), the *social dimensions of annotators* like their demographic (Sap et al., 2019; Larimore et al., 2021; Sap et al., 2022) and cultural backgrounds (Hershovich et al., 2022), often discussed more generally as different perceptions in *data perspectivism* (Basile et al., 2021a; Wich et al., 2021). Moreover, there is work that acknowledges that **multiple plausible answers** are correct, such as works on the *collective human opinion* (Nie et al., 2020) influenced by seminal work that looks at the *human judgement distribution* (Pavlick and Kwiatkowski, 2019) who found plausible variation in at least 20% of their data. Earlier work on

veridicality also made this point (de Marneffe et al., 2012). The fact that multiple plausible annotations exist has also been put forward as *a range of acceptable annotations* (Palomaki et al., 2018). The known variation in annotation for subjective tasks is at least a decade old (Alm, 2011). They suggest that in the absence of a real ‘ground truth’, acceptability may be a more useful concept than ‘right’ and ‘wrong’. Capturing the HLV, instead of the global majority, aligns with this viewpoint.

Open issues and our suggestions To make progress, we need to i) collect and release annotator-level (un-aggregated) labels, ii) document dataset creation, and iii) include as much meta-data as possible. In particular, we urge the community to release annotator-level (un-aggregated) labels—even if only for a small subset of the data—and thus we echo Basile et al. (2021b) and Prabhakaran et al. (2021) (also in Denton et al. (2021)) who independently raised this point as well.

As a concrete starting point, we provide a comprehensive overview of existing datasets with multiple annotations in the appendix, which we release as a github repository to encourage uptake. Moreover, if possible to release responsibly, besides making data statements of datasets available (Bender and Friedman, 2018), we encourage the community to include annotator-level background information (Prabhakaran et al., 2021) and document the annotation process (Geiger et al., 2020). In general, we believe there is high value in releasing any meta-data available (ideally on the instance level, e.g. source, time of document, annotator ids, annotation completion time etc). For example, in a recent study we created a new relation extraction corpus with instance-level flags of annotator uncertainty proving valuable for evaluation (Bassignana and Plank, 2022). Similarly, we asked the annotator to provide free-text rationales of relations, which recently was also put forward in Borin (2022), referring to earlier work on collecting annotator rationales during annotation (McDonnell et al., 2016).

We believe that the more, richer datasets become available, the more insights can be generated into the capabilities of models and their limitations. New algorithms may emerge capable of learning from fewer but richer sources. On a related line, collecting multiple annotations calls for research in estimating data quality and revisiting agreement measures; e.g., new measures for multiple-labels were recently proposed (Marchal et al., 2022).

3 Modeling and Human Label Variation

There is a growing literature on methods on how to deal with HLV in learning. We categorize them into two camps: those that resolve variation, and those that embrace it. We will draw connections to surveys and the emerging literature, and discuss adoption of methods as well as gaps.

The first big camp of research aims at **resolving human label variation** and includes: 1) Aggregation and 2) Filtering. It considers HLV as “problematic” or “noisy”. Consequently, a single (aggregated) label is obtained with presumably high agreement as the ground truth. *Aggregation* is performed via majority voting or probabilistic aggregation methods, see Paun et al. (2022) for a survey and seminal works (Dawid and Skene, 1979; Qing et al., 2014; Artstein and Poesio, 2008). Aggregation is still the most widely-adopted solution for the problem today. However, aggregation by definition allows only *one belief/label/category*. This is very limiting, as often it is not just about disagreement or matter of subjectivity, but multiple options being plausible. *Filtering* methods are advocated by some with the idea to remove data instances with low agreement (Reidsma and Carletta, 2008; Reidsma and op den Akker, 2008; Beigman Klebanov et al., 2008; Beigman and Beigman-Klebanov, 2009). However, only using high-agreement instances can yield worse performance (Jamison and Gurevych, 2015) and it wastes data.

The second camp of research instead aims at **embracing human label variation**. Two broad directions include: 3) Learning from un-aggregated labels (directly), or 4) Enriching gold with human label variation. With regard to *learning from un-aggregated labels*, methods of varying complexity exist, from model-agnostic methods such as repeated labeling (Sheng et al., 2008) used by e.g. de Marneffe et al. (2012), to architecture-specific choices, e.g., adding a crowd layer (Rodrigues and Pereira, 2018), learning from soft labels (Peterson et al., 2019) and more; see the survey of Uma et al. (2021b). So far learning from un-aggregated labels directly has shown greater promise in classification tasks in CV than in NLP (Uma et al., 2021b) (evidence is scarce, see open issues). Within NLP, a more studied direction is currently *to enrich the gold label with human label variation*, i.e., to learn from both the gold and the un-aggregated labels. Methods in this category can be seen as part of the broader set of well-

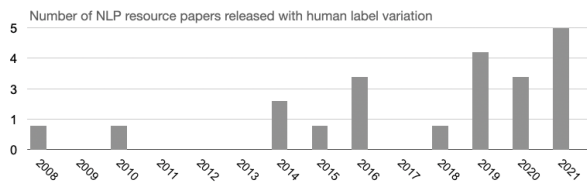


Figure 3: NLP Resource papers per publication year, counting publicly-available datasets released with human label variation (multiple annotator-labels per instance), cf. details in Table 1 in the Appendix.

known regularization methods in ML, and for NLP include e.g., cost-sensitive loss weighting (Plank et al., 2014a), variants of multi-task learning (Cohn and Specia, 2013; Fornaciari et al., 2021; Davani et al., 2022), or sequential fine-tuning (Lalor et al., 2017). These methods further differ in how they use un-aggregated labels, i.e., as confusion matrices estimated from a small sample (Plank et al., 2014a), as annotator-level auxiliary tasks requiring the full data with multiple labels (Cohn and Specia, 2013; Davani et al., 2022), or as single “soft-label” auxiliary task that captures the per-instance human label distribution (Fornaciari et al., 2021).

Open issues and our suggestions Undoubtedly, there is increasing interest in studying methods to learn with human label variation (see Figure 3 for our analysis of research papers). However, existing research is fragmented across (sub-)disciplines. We identify at least three diverse areas within NLP, with little to no overlap (as shown in Table 1 in the Appendix), focusing respectively on: subjectivity (Basile et al., 2021a) (`pdai.info`, SemEval 23), natural language inference (NLI) (Pavlick and Kwiatkowski, 2019; Nie et al., 2020), and both NLP and CV (JAIR & SemEval 21). To the best of our knowledge, only the latter work and shared task so far bridges across disciplines (Uma et al., 2021b,a). Still, they focus on complementary NLP tasks to the two previous initiatives. It is thus an open issue to see whether tasks might need to have specific properties to be suitable for one kind of method over another. A comprehensive evaluation is lacking. Studying transferability of methods across problems is another interesting open issue.

Learning from HLV heavily depends on data labeled with multiple annotators. In some settings, it might be difficult to obtain sizeable amounts of such data (however, as seen in Section 2, more datasets are emerging). Regarding learning, Lalor et al. (2017) find that even small amounts of data

can be helpful in a sequential fine-tuning setup, as also early work indicates (Plank et al., 2014a). An open challenge is to find the right balance between the amount of data collected and the number of annotators. Overall, we hypothesize that the richness of information captured by human label variation has the potential to reduce data size requirements (possibly fewer instances but with more information captured in the human label distribution). It remains an open issue to connect with emerging works on learning with different amounts of annotation (Zhang et al., 2021), which can also lead to novel architectures.

A related important challenge is to tease apart errors from signal (e.g. Reidsma and Carletta, 2008; Gordon et al., 2021). Work on annotation error detection exists, cf. the very recent survey by Klie et al. (2022) or Zhang and de Marneffe (2021). It is though largely overlooked. This calls further for theoretical work on the notion of an what constitutes an error versus a hard case (Manning, 2011; Webber and Joshi, 2012; Plank et al., 2014b). This bears connections to emerging work in HCI, in particular social computing (Gordon et al., 2021, 2022), who look at the perception of system errors by humans, see also Section 4, and earlier work in HCI on crowdsourcing that allows for some errors (Krishna et al., 2016).

While embracing human label variation helps to regularize learning, the connection to a broader range of ML methods such as noise labeling or calibration remains highly relevant and a source of further inspiration (Goldberger and Ben-Reuven, 2016; Han et al., 2018b,a; Meister et al., 2020). There are some initial studies that compare human disagreement with model confidence (Davani et al., 2022). Overall, interest in calibration methods (Naeini et al., 2015; Guo et al., 2017) is increasing (Desai and Durrett, 2020; Kong et al., 2020; Jiang et al., 2021) to counter overconfidence of neural classifiers (Meister et al., 2020). In contemporary work to this, we show that measuring calibration to human majority given inherent disagreements is theoretically and empirically problematic (Baan et al., 2022). As a first step, we propose instance-level measures of calibration that better capture the human label distribution. In future, it remains to be seen how to best use human label variation to make systems more trustworthy.

Finally, there is relevant interesting work that more deeply looks at data during learning. In NLP,

recent seminal work by Swayamdipta et al. (2020) proposes *data maps* to investigate the behavior of a model on individual instances during training (training dynamics). They show that training a system on *ambiguous* instances identified via data maps helps to generalize better in out-of-distribution evaluation (Swayamdipta et al., 2020). Building on top of this work, Zhang and Plank (2021) show that the instances at the boundary of hard and ambiguous cases derived from small data maps aids active learning. This is further evidence that human uncertainty in labeling is beneficial for learning. It remains to be seen whether training dynamics can yield novel architectures for learning from HLV.

4 Evaluation and Human Label Variation

Evaluation is of critical importance in empirical research fields such as ML, NLP and CV. It helps to choose one system over another, and to measure progress. However, current evaluation practices typically use accuracy against a gold standard. In many tasks this common practice is severely flawed. It obfuscates the truth about the state of ML models. It leaves a large gap between in-vitro and in-vivo evaluation. HCI research has shown that metrics are not aligned with reality; audits of algorithms' performance have uncovered very poor results in practice, and that this disconnect is indicative of a larger disconnect on how ML and HCI researchers evaluate their work (Gordon et al., 2021). We believe this is an important take-away for NLP. We too often focus on single metrics, single components of the pipeline, in other words, on myopic *in-vitro* experimentation.

Open issues and our suggestions Despite the increasing body of literature on methods for learning with HLV, a majority of the papers introducing new methods strikingly evaluate against *hard labels* (gold labels) (e.g. Rodrigues and Pereira, 2018; Fornciari et al., 2021). If we want to take human label variation seriously, we need to shift our attention to evaluation that goes beyond hard labels (accuracy). As accuracy of all models can be high (at times), looking at only one metric (and, in fact a single—argmax—prediction) gives no indication on how reasonable a model is, yet alone how confident and trustworthy it is.

Research in ML, CV and NLP has started to incentivize *hard and soft label* evaluation. Soft labels compare the human label distribution to model outputs. Proposed soft metrics include: *cross entropy*,

to capture how well the model captures humans' assessment not just of the top label, which is used in both CV (Peterson et al., 2019) and NLP (Pavlick and Kwiatkowski, 2019); *entropy correlation* proposed by Uma et al. (2020), to compute Pearson's correlation between instance-level entropy scores of human soft labels and model predictions; *Kullback-Leibler divergence*-based evaluation (Nie et al., 2020) (either KL or Jensen-Shannon). Others instead started to evaluate against *individual annotators* (Resnick et al., 2021; Davani et al., 2022), measure F1 scores against *data splits by different annotator agreement levels* (Leonardelli et al., 2021; Damgaard et al., 2021), data splits based on *annotator clustering* (Basile et al., 2021a), data splits based on *item difficulty* based on entropy of the label distribution and semantic distance (Jolly et al., 2021), and data splits based on *annotator uncertainty flags* (Bassignana and Plank, 2022). Analogously as in Section 3, it is an open issue to see whether tasks might need to have specific properties to be more suitable for one kind of evaluation over another. In general, we need better evaluation practices (besides soft and hard evaluation), particularly in light of the complexity of human label variation—and the reasons it arises, which might be due to uncertainty, background, task complexity, intra-coder reliability etc; see Basile et al. (2021b) and in particular Jiang and de Marneffe (2022) for a discussion on disagreement sources; the latter recently developed a taxonomy for disagreement in natural language inference data.

5 Conclusions

In this paper, we outline that human label variation impacts all steps of the traditional ML pipeline, and is an opportunity, not a problem. To move forward, we argue for a more comprehensive treatment of HLV, which considers all steps, to enable innovation: data, modeling and evaluation. To do so, and truly move beyond the current in-vitro setups, we need an open, interdisciplinary discussion. We hope to contribute to this discussion, and stipulate research with the released repository: <https://github.com/mainlp/awesome-human-label-variation>.¹

¹The repository contains the datasets in Appendix 1 as a starting point. This is, to the best of our knowledge, the most comprehensive list of datasets with un-aggregated labels available today. We encourage readers to contribute. They are further invited to join the SemEval 2023 shared task LeWiDi: <https://le-wi-di.github.io/>

Limitations

This position paper tries to be succinct while aiming at synthesizing a very broad notion—human label variation—that affects all steps dealing with learning from annotated data. Therefore, this position paper is necessarily incomplete, as is the dataset repository that is provided. However, we hope that the repository and paper will lead to an open discussion and community uptake, as this is a big open issue and necessitates a broader, interdisciplinary treatment.

Ethics Statement

Modeling human label variation is connected to social bias, as annotator backgrounds influence annotations and consequently both machine learning and evaluation. Therefore it is important to be aware of possible social implications of some of the technologies discussed here. Inevitably there is potential for dual use, as amplifying the voice of some might harm others. However, there are social opportunities, as modeling human label variation allows to include the voices of more groups, and even the very underrepresented. In a world where the majority view dominates, these would otherwise be left behind.

Acknowledgements

We thank the reviewers for their feedback. Special thanks to Bonnie Webber, Massimo Poesio and Raffaella Bernardi for invaluable feedback on drafts of this paper and discussions on this topic, in parts at the Insights workshop@ACL 2022. Thanks to members of the NLPnorth & MaiNLP lab for feedback on this paper. BP is supported by ERC Consolidator Grant DIALECT 101043235.

References

- Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.
- Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.
- Dina Almanea and Massimo Poesio. 2022. The ARMIS dataset of misogyny in arabic tweets. In *Proc. of LREC*.
- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Empirical Methods of Natural Language Processing: EMNLP 2022*. Association for Computational Linguistics.
- Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021a. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021b. We need to consider disagreement in evaluation. In *1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.
- Elisa Bassignana and Barbara Plank. 2022. CrossRE: A Cross-Domain Dataset for Relation Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.
- Eyal Beigman and Beata Beigman-Klebanov. 2009. [Learning with annotation noise](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, Suntec, Singapore. Association for Computational Linguistics.
- Beata Beigman Klebanov, Eyal Beigman, and Danie Diermaier. 2008. Analyzing disagreements. In *Proceedings of the Coling 2008 workshop on Human Judgements in Computational Linguistics*, pages 2–7.
- Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. 2016. [Anchoring and agreement in syntactic annotations](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2224, Austin, Texas. Association for Computational Linguistics.
- Lars Borin. 2022. All that glitters... interannotator agreement in natural language processing. *Morfologi, målstrev og maskinar – Trond Trosterud {fyller ||tytt||deavd||turns}60!*, 46(1).

- Christopher Bryant and Hwee Tou Ng. 2015. [How far are we from fully automatic high quality grammatical error correction?](#) In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.
- Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. [ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI.](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Veronika Cheplygina and Josien PW Pluim. 2018. Crowd disagreement about medical images is informative. In *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis*, pages 105–111. Springer.
- Trevor Cohn and Lucia Specia. 2013. [Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation.](#) In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. Association for Computational Linguistics.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.
- Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. [“I’ll be there for you”: The one with understanding indirect answers.](#) In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Debopam Das, Manfred Stede, and Maite Taboada. 2017. [The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis.](#) In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 11–19, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.
- Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. [Did it happen? the pragmatic complexity of veridicality assessment.](#) *Computational Linguistics*, 38(2):301–333.
- Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
- Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. In *NeurIPS workshop on Data-Centric AI*.
- Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. [Broad Twitter corpus: A diverse named entity recognition resource.](#) In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.
- Shrey Desai and Greg Durrett. 2020. [Calibration of pre-trained transformers.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.
- Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. [A computational exploration of pejorative language in social media.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. [Crowdsourcing semantic label propagation in relation classification.](#) In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium. Association for Computational Linguistics.
- Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. [Did they answer? subjective acts and intents in conversational discourse.](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

- Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. [Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.
- R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336.
- Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer. *ICLR 2017*.
- Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 1321–1330. JMLR.org.
- Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. 2018a. Masking: A new perspective of noisy supervision. In *Advances in Neural Information Processing Systems*, pages 5836–5846.
- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018b. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsourcing annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1081–1097.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Shailza Jolly, Sandro Pezzelle, and Moin Nabi. 2021. [EaSe: A diagnostic tool for VQA based on answer diversity](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2407–2414, Online. Association for Computational Linguistics.
- Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multi-task deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.
- Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2022. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*.
- Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. [Calibrated language model fine-tuning for in- and out-of-distribution data](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. 2016. Embracing error to enable rapid crowdsourcing. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3167–3179.
- John P Lalor, Hao Wu, and Hong Yu. 2017. Soft label memorization-generalization for natural language inference. *arXiv preprint arXiv:1702.08563*.

- Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tong Liu, Christopher Homan, Cecilia Ovesdotter Alm, Megan Lytle, Ann Marie White, and Henry Kautz. 2016. Understanding discourse on work and job-related well-being in public social media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1053, Berlin, Germany. Association for Computational Linguistics.
- Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.
- Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Héctor Martínez Alonso, Anders Johannsen, and Barbara Plank. 2016. Supersense tagging with inter-annotator disagreement. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 43–48, Berlin, Germany. Association for Computational Linguistics.
- Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, pages 139–148.
- Clara Meister, Elizabeth Salesky, and Ryan Cotterell. 2020. Generalized entropy regularization or: There’s nothing special about label smoothing. *arXiv preprint arXiv:2005.00820*.
- Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, page 2901–2907. AAAI Press.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *SAD/CrowdBias@ HCOMP*, pages 19–31.
- Rebecca J Passonneau, Ansa SALLEB-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*.
- Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. Statistical methods for annotation analysis. *Synthesis Lectures on Human Language Technologies*, 15(1):1–217.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. Gcdt: A chinese rst treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.
- Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751. Association for Computational Linguistics.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.
- Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.
- Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

- 1 (*Long and Short Papers*), pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. [On releasing annotator-level labels and information in datasets](#). In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O’Reilly Media.
- Ciyang Qing, Ulle Endriss, Raquel Fernández, and Justin Kruger. 2014. Empirical analysis of aggregation methods for collective annotation. In *COLING*.
- Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.
- Dennis Reidsma and Rieks op den Akker. 2008. [Exploiting ‘subjective’ annotations](#). In *Proc. of the Workshop on Human Judgments in Computational Linguistics*, pages 8–16.
- Paul Resnick, Yuqing Kong, Grant Schoenebeck, and Tim Weninger. 2021. Survey equivalence: A procedure for measuring classifier accuracy against human labels. *arXiv preprint arXiv:2106.01254*.
- Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- David Schlangen. 2021. [Targeting the benchmark: On methodology in current natural language processing research](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.
- Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. [Get another label? improving data quality and data mining using multiple, noisy labelers](#). In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’08*, page 614–622, New York, NY, USA. Association for Computing Machinery.
- Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. [Predicting humorousness and metaphor novelty with Gaussian process preference learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.
- Rion Snow, Brendan O’connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.
- Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. [Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. [Dataset cartography: Mapping and diagnosing datasets with training dynamics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.
- Alexandra Uma, Tommaso Fornaciari, Anca Dumitrescu, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. [SemEval-2021 task 12: Learning with disagreements](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.
- Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft-loss functions. In *Proc. of HCOMP*.
- Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

- Bonnie Webber and Aravind Joshi. 2012. [Discourse structure and computation: Past, present and future](#). In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju Island, Korea. Association for Computational Linguistics.
- Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. 2021. [Investigating annotator bias in abusive language datasets](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1515–1525, Held Online. INCOMA Ltd.
- Daniel Zeman. 2010. Hard problems of tagset conversion. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 181–185.
- Mike Zhang and Barbara Plank. 2021. [Cartography active learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. *arXiv preprint arXiv:2109.04408*.
- Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. [Identifying inherent disagreement in natural language inference](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

A Datasets with Multiple Annotations

Field	Reference	Name/Description	URL	Jair	PDai	SemEval21	TACL22	SemEval23
NLP	(Passonneau et al., 2010)	Word sense disambiguation (WSD)	https://anc.org/					
	(Plank et al., 2014a)	POS tagging (500 tweets from Lowlands) and Gimpel-POS dataset	https://bitbucket.org/lowlands/costsensitive-data/ and https://zenodo.org/record/5130737	✓		✓		
	(Derczynski et al., 2016)	NER Broad Twitter dataset	https://github.com/GateNLP/broad_twitter_corpus		✓			
	(Rodrigues and Pereira, 2018)	NER dataset, re-annotated sample of CoNLL 2003	http://fprodrigues.com/publications/deep-crowds/					
	(Martínez Alonso et al., 2016)	Supersense tagging	https://github.com/coastalcp/semdux					
	(Berzak et al., 2016)	Dependency Parsing, WSJ-23, 4 annotators	https://people.csail.mit.edu/berzak/agreement/					
	(Peng et al., 2022)	GCDT, Mandarin Chinese discourse treebank	https://github.com/logan-siyao-peng/GCDT/tree/main/data					
	(Bryant and Ng, 2015)	Grammatical error correction	http://www.comp.nus.edu.sg/~nlp/sw/10gec_annotations.zip					
	(Poesio et al., 2019)	PD (Phrase Detectives dataset): Anaphora and Information Status Classification	https://github.com/dali-ambiguity/Phrase-Detectives-Corpus-2.1.4	✓		✓		
	(Dumitrache et al., 2018)	Medical Relation Extraction (MRE)	https://github.com/CrowdTruth/Open-Domain-Relation-Extraction	✓				
	(Bassignana and Plank, 2022)	CrossRE, relation extraction, small doubly-annotated subset	https://github.com/mainlp/CrossRE					
	(Dumitrache et al., 2019)	Frame Disambiguation	https://github.com/CrowdTruth/FrameDisambiguation					
	(Snow et al., 2008)	RTE (recognizing textual entailment; 800 hypothesis-premise pairs) collected by (Dagan et al., 2005), re-annotated; includes further datasets on temporal ordering, WSD, word similarity and affective text	https://sites.google.com/site/nlpannotations/	✓				
	(Pavlick and Kwiatkowski, 2019)	NLI (natural language inference) inherent disagreement dataset, approx. 500 RTE instances from (Dagan et al., 2005) re-annotated by 50 annotators	https://github.com/epavlick/NLI-variation-data					
	(Nie et al., 2020)	ChaosNLI, large NLI dataset re-annotated by 100 annotators	https://github.com/easonnie/ChaosNLI					
	(Demszky et al., 2020)	GoEmotions: reddit comments annotated for 27 emotion categories or neutral	https://github.com/google-research/google-research/tree/master/goemotions				✓	
	(Ferracane et al., 2021)	Subjective discourse: conversation acts and intents	https://github.com/elisaf/subjective_discourse					
	(Damgaard et al., 2021)	Understanding indirect answers to polar questions	https://github.com/friendsQIA/Friends_QIA					
	(de Marneffe et al., 2019)	CommitmentBank: 8 annotations indicating the extent to which the speakers are committed to the truth of the embedded clause	https://github.com/mcdm/CommitmentBank					
	(Kennedy et al., 2020)	Hate speech detection	https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech		✓		✓	
	(Dinu et al., 2021)	Pejorative words dataset	https://nlp.unibuc.ro/resources or http://pdai.info/		✓			
	(Leonardelli et al., 2021)	MultiDomain Agreement, Offensive language detection on Twitter, 5 offensive/non-offensive labels; also part of LeWiDi SemEval23	https://github.com/dhfbk/annotators-agreement-dataset/		✓			✓
	(Cercas Curry et al., 2021)	ConvAbuse, abusive language towards three conversational AI systems; also part of LeWiDi SemEval23	https://github.com/amandacurry/convabuse		✓			✓
(Liu et al., 2016)	Work and Well-being Job-related Tweets, 5 annotators	https://github.com/Homan-Lab/pid1_data		✓				
(Simpson et al., 2019)	Humour: pairwise funniness judgements	https://zenodo.org/record/5130737			✓			
(Akhtar et al., 2021)	HS-brexite; New LeWiDi-23 shared tast dataset on Abusive Language on Brexit and annotated for hate speech (HS), aggressiveness and offensiveness, 6 annotators	https://le-wi-di.github.io/		✓			✓	
(Almanea and Poesio, 2022)	ArMIS; New LeWiDi-23 shared tast dataset on Arabic tweets annotated for misogyny detection	https://le-wi-di.github.io/					✓	
CV	(Rodrigues and Pereira, 2018)	LabelMe: Image classification dataset with 8 categories, re-annotated	http://fprodrigues.com/publications/deep-crowds/	✓		✓		
	(Peterson et al., 2019)	Cifar10H: Image classification with 10 categories, re-annotated	https://github.com/jcpeterson/cifar-10h	✓		✓		
	(Cheplygina and Plum, 2018)	Medical lesion classification challenge, 6 annotators each	https://figshare.com/s/5ebbc14647b66286544					

Table 1: Overview of publicly-available datasets with HLV data (see repository for updates and to contribute: <https://github.com/mainlp/awesome-human-label-variation>). ✓: whether the source was used in broader empirical evaluations, e.g., the JAIR survey on learning from disagreement (Uma et al., 2021b), is listed on pdai.info (Basile et al., 2021a) (as of June, 2022), is part of the SemEval 2021 task on learning from disagreement (Uma et al., 2021a), is used in a TACL paper on learning beyond majority vote (Davani et al., 2022), is used in the SemEval 2023 shared task on Learning With Disagreement LeWiDi <https://le-wi-di.github.io/>.